Pour une micro-diachronie de l'oral : le corpus ESLO-MD

Lotfi Abouda[1], *Marie* Skrovec[2]
Laboratoire Ligérien de Linguistique (LLL-UMR 7270), Université d'Orléans

lotfi.abouda@univ-orleans.fr

Résumé. Extrait des Enquêtes Socio-Linguistiques à Orléans, corpus oral constitué en deux temps à 40 ans d'intervalle (ESLO 1968-1971, ESLO2 2008-), ESLO-MD (ESLO Micro-Diachronie) est un corpus oral échantillonné d'un million de mots, quantitativement et qualitativement équilibré entre ses deux parts micro-diachroniques. Constitué initialement en 2014 pour les besoins d'une étude particulière, le corpus ESLO-MD a montré, au fil des ans et des recherches, une certaine efficacité à éclairer sous un jour nouveau un certain nombre de phénomènes linguistiques. Par la prise en compte d'une dimension micro-diachronique inédite, ESLO-MD permet en effet utilement de compléter l'éventail des données disponibles pour l'étude de la variation et du changement en français oral hexagonal contemporain. Cette contribution, qui entend présenter le travail de constitution de ce corpus, parallèlement à sa mise à disposition pour la communauté scientifique, expose les critères de sélection des données ainsi que les modalités de leur diffusion, dans la perspective d'un accès facilité à des données sources susceptibles d'être réexaminées ou réexploitées par d'autres membres de la communauté.

Abstract. Extract from the Enquêtes Socio-Linguistiques à Orléans, oral corpus constituted in two stages at 40 years intervals (ESLO 1968-1971, ESLO2 2008-), ESLO-MD (ESLO Micro-Diachronie) is an oral corpus sampled of one million words, quantitatively and qualitatively balanced between its two micro-diachronic parts. Originally created in 2014 for the purpose of a particular study, the ESLO-MD corpus has shown, over the years and through research, a certain effectiveness to shed new light on a certain number of linguistic phenomena. By taking into account an unprecedented micro-diachronic dimension, ESLO-MD makes it possible to usefully complete the range of data available for the study of variation and change in contemporary oral French. This contribution, which aims to present the work of compiling this corpus, in parallel with its availability to the scientific community, sets out the criteria for selecting data and the modalities of their dissemination, with a view to facilitating access to source data that can be reviewed or reused by other members of the community.

1 Introduction

Les Enquêtes Socio-Linguistiques à Orléans (désormais ESLO[3] possèdent une triple caractéristique qui leur permet d'occuper une place singulière dans le paysage des bases de données sur le français oral. En plus de contenir des données quantitativement significatives - actuellement environ 6,5 millions de mots -, qui en font l'un des plus vastes corpus oraux disponibles, ESLO est composé de données langagières dites "situées", i.e. dont la production est contextualisée sous forme de métadonnées, qui renseignent sur la date et le lieu de la collecte, le type de l'interaction (entretiens semi-guidés, repas de famille ou entre amis, conférences universitaires, etc.) et sur le profil des locuteurs, répartis en tranches d'âge, de sexes et de catégories socio-professionnelles. Enfin, ce corpus a la particularité d'avoir été constitué en deux temps, à 40 ans d'intervalle, puisqu'à une première campagne de collecte réalisée de 1968 à 1971 par des universitaires britanniques (ESLO1) a succédé, depuis 2010, une deuxième campagne cherchant à collecter des données en partie comparables (ESLO2), ce qui donne aux ESLOs une importance particulière dans le champ d'étude émergent de la micro-diachronie.

C'est la conjonction de ces trois facteurs qui confère sa singularité à ce corpus qui vient ainsi compléter l'éventail des données disponibles pour l'étude de la variation et du changement en français. En effet, si des corpus oraux existent en grand nombre depuis quelques années (CLAPI, TCOF, PFC, CFPP...), l'approche entreprise y est généralement résolument synchronique, tandis que le champ des études en micro-diachronie, s'il permet de considérer différents empans diachroniques, se fonde essentiellement sur des données écrites, plus faciles d'accès et plus nombreuses (cf. Siouffi, 2012). C'est précisément pour pallier cette insuffisance que la plupart des études disponibles sur le changement à l'oral adoptent une démarche d'observation du changement dit "en temps apparent" (cf. Fleury & Branca 2010). Le peu d'études en temps réel (ou, plus précisément mixtes, combinant temps réel et temps apparent) qui existent dans le domaine francophone, en plus d'avoir des caractéristiques distinctes d'ESLO (il s'agit essentiellement de données longitudinales recueillies en deux ou trois fois auprès des mêmes locuteurs avec un empan maximal de 25 ans), ne portent que sur le français québécois (voir notamment Thibault & Vincent 1990 et Blondeau & al. 2002). Il apparaît donc clairement que, dans le domaine de la documentation du français oral hexagonal, le recul diachronique de 40 ans que permet ESLO est fondamentalement novateur.

Conçu dès le départ « comme un réservoir qui doit permettre à un chercheur de construire son propre corpus » (Abouda & Baude, 2007 : 165), ESLO offre au chercheur la possibilité de constituer, en fonction de l'hypothèse projetée, des collections particulières de données, résultant d'une combinaison précise, explicitable et justifiable, de métadonnées paramétrables sur les plans diachronique, diastratique et diaphasique. C'est ainsi que, étudiant un phénomène linguistique particulièrement complexe et discuté, la distribution futur simple-futur périphrastique[4], nous avons entrepris la constitution d'un sous-corpus, ESLO-MD (ESLO Micro-Diachronie), composé de deux parts diachroniquement différenciées (ESLO1-ESLO2) mais qualitativement et quantitativement comparables. Sélectionné d'une manière rigoureuse, documenté et enrichi par une annotation fine, ce corpus doit à notre avis être partagé, non seulement pour faciliter les procédures de vérification de nos propres analyses, étape selon nous désormais incontournable d'une démarche scientifique, mais aussi parce qu'il nous semble particulièrement bien placé pour alimenter la réflexion sur l'articulation entre variation et changement, et pourrait donc servir de base de données à de nouvelles recherches, portant aussi bien sur des phénomènes linguistiques et leurs éventuelles évolutions micro-diachroniques que sur l'usage de ces phénomènes et leurs éventuelles évolutions internes et/ou externes.

L'objet de cet article est la présentation problématisée du corpus ESLO-MD, parallèlement à sa mise à disposition pour la communauté scientifique. Il s'agira, d'une part, de contextualiser sa constitution, nécessaire à son exploitation pour d'autres études linguistiques, et de décrire notre méthodologie d'annotation et d'enrichissement des données, et, d'autre part, de préciser les modalités et plus-values de sa mise à disposition. Enjeu important dans le champ de la linguistique sur corpus oraux, la mise à disposition s'inscrit dans les pratiques actuelles de transparence par le partage de données primaires et secondaires de différents types (transcription, métadonnées, annotations). En plus de permettre la vérification, elle rend les données réexploitables pour d'autres objets, et favorise ainsi la cumulativité des connaissances en sciences sociales (Pumain 2005).

2 Critères de sélection du corpus ESLO-MD

Par opposition à une époque pas si éloignée où les données, parce qu'elles étaient difficiles à rassembler, pouvaient être relativement bien conçues, l'ère numérique, en rendant possible un accès désormais (trop) aisé à des ressources électroniques disponibles en grand nombre, voit apparaître le risque nouveau d'une "confusion dans la profusion" (Habert, 2005 : 13). Pris au piège du "big is beautiful" (Svartvik, 1992 : 10), le chercheur peut être tenté de constituer des corpus toujours plus grands, au détriment de leur composition interne, comme si la probité des résultats d'une recherche ne dépendait que de la taille du corpus exploré, et si peu de sa qualité et de sa cohérence interne (voir notamment Habert 2000 & 2005).

L'objectif de comparabilité avec ESLO1, qui a initialement présidé à la constitution d'ESLO2, a dû céder face à l'impératif d'adaptation aux nouvelles exigences scientifiques et possibilités techniques contemporaines. ESLO2 sera finalement, par bien des aspects, différent de son aîné. Or la perspective micro-diachronique de nos recherches nécessitait le recours à des données offrant le meilleur équilibre quantitatif et

qualitatif possible entre ses deux parts micro-diachroniques. En plus de cette contrainte, le corpus à constituer devait respecter une seconde condition non moins importante : sa taille ne devait pas rendre impraticable la tâche, particulièrement chronophage, de son enrichissement sémantique.
Ces considérations nous ont amenés à constituer un corpus d'environ 1 million de mots (soit un peu plus de 80 heures d'enregistrement).

Tableau 1. Caractéristiques quantitatives du corpus ESLO-MD.

	ESLO1	**ESLO2**	**Total**
Durée (en min)	2430	2421	4851 (80h et 51 min)
Nombre de mots	453 298	521 931	975 229

2.1 Equilibrage diachronique, diaphasique et diastratique

Les deux extraits ESLO1/ESLO2 équilibrés quantitativement devaient offrir le maximum d'équilibrage qualitatif possible.

Sur le plan diaphasique, il était nécessaire que les deux extraits comparés contiennent les mêmes types d'interaction. Or, ainsi que nous pouvons le constater dans le tableau ci-dessous, la zone d'intersection diaphasique (en grisé) ne comporte que trois genres interactionnels : les entretiens, les conférences et les repas en famille ou entre amis.

Tableau 2. Modules ESLO1/ESLO2

ESLO1	**ESLO2**
Appel téléphonique	24H[5]
Conférences	Boulangerie
Consultation CMPP[6]	Cinéma
Contact	Conférences
Divers	Diachronie
Entretiens	Discours
Interview de personnalités	Ecole
Magasin	Entretiens
Marché	Entretiens chercheurs
Repas	Entretiens jeunes
Visites	Itinéraires
	Livres pour Enfants
	Média

L'exigence de comparabilité diaphasique globale entre les deux extraits ESLO1/2 n'a pas en revanche permis d'assurer un équilibre diaphasique à l'intérieur de chacun d'entre eux. En effet, ainsi que nous le verrons ci-dessous, l'exigence parallèle de comparabilité diastratique nécessitait un nombre minimum d'entretiens (pour avoir un représentant de chacune des sous-catégories retenues) face auxquels on ne disposait pas de suffisamment de données dans les repas et/ou les conférences dans l'un ou l'autre des deux ESLO. Dans un corpus disponible majoritairement constitué d'entretiens, nous avons tout de même réussi à intégrer, à hauteur de 20%, deux genres interactionnels de « contrôle », supposés se situer de part et d'autre d'une échelle diaphasique, i.e. les repas (en famille ou entre amis) et les conférences.

Ces choix sont résumés dans le tableau ci-dessous :

Tableau 3. Répartition diaphasique dans ESLO-MD

		ESLO1	**ESLO2**	**TOTALE**
Durée (en min)	Conférences	192	186	378
	Repas	196	201	397
	Entretiens	2042	2034	4076
	Totale	2430	2421	4851

À l'intérieur de cette composition, s'est parallèlement posée la question de la sélection des enregistrements, notamment pour les entretiens. S'agissant d'un corpus de données situées, nous avons essayé de garantir au mieux un équilibrage diastratique (entre ESLO1 et ESLO2 et à l'intérieur de chacun d'entre eux), en termes de profil de locuteurs, choisis selon les variables de sexe (H/F), d'âge (3 tranches) et de catégorie socio-professionnelle (5 classes). Concrètement, notre objectif était d'obtenir, pour chacun des deux extraits micro-diachroniques de notre corpus (ESLO1/2), 1 représentant homme et 1 représentant femme pour chacune des 5 CSP, dans chacune des 3 tranches d'âge, soit en tout 60 locuteurs.

Pour opérer une telle sélection en fonction de ces trois variables, nous avons pris quelques décisions et dû consentir à quelques aménagements.

La première décision a été de classer les locuteurs en trois tranches d'âge : 15-35, 35-60 et plus de 60 ans. Les données disponibles et la nécessité d'avoir un représentant de chaque sexe pour chaque CSP ne pouvaient pas permettre une répartition plus fine en tranches d'âge. Lorsque, à l'intérieur d'une tranche d'âge, les autres variables nous laissaient le choix entre plusieurs témoins, nous avons privilégié ceux dont l'âge était le plus éloigné des tranches voisines.

Concernant la variable de sexe, la combinaison des différents critères n'a pas toujours permis d'atteindre notre objectif initial d'avoir, globalement et à l'intérieur de chaque classe d'âge et de chaque CSP, autant d'hommes que de femmes. Donnant la priorité aux deux autres critères (âge et CSP), nous avons dû accepter que cette parité, lorsqu'elle n'est pas atteinte dans chaque combinaison tranche d'âge-CSP, puisse l'être au niveau d'un couple de CSP contigües. Cet aménagement a permis de compenser deux des trois cas d'absence de

parité. Il n'a pas été en revanche possible de compenser le troisième cas de déséquilibre en sexe, puisque, afin de nous approcher le plus possible d'un équilibre quantitatif global entre ESLO1 et ESLO2 (mesuré en minutes et non en nombre d'entretiens), nous avons intégré un entretien supplémentaire dans le sous-corpus ESLO2, ce qui ne pouvait qu'affecter l'équilibre en termes de sexe et de CSP.

Lorsque la conjonction de ces trois variables principales nous laissait le choix entre plusieurs entretiens, nous avons fait appel à deux autres critères de sélection. Le premier a consisté à privilégier les locuteurs « DIA » (locuteurs interrogés successivement dans ESLO1 et ESLO2, soit 13 locuteurs). Le second critère a été de choisir l'entretien dont la durée est la plus proche de 60 minutes.

2.2 Variables sociologiques : adaptation de l'échelle AM

Afin de garantir une certaine comparabilité entre les deux parts micro-diachroniques de notre corpus, les compromis les plus importants, et sans doute les plus discutables, que nous avons dû consentir ont concerné les variables sociologiques, tant il était hasardeux d'établir des parallèles entre des locuteurs d'ESLO1 et d'ESLO2, classés sociologiquement selon des critères différents (et pas toujours précisément explicités).

Dans l'enquête initiale, les initiateurs d'ESLO1 ont fait preuve, ainsi que le souligne Bergounioux & al. (1992) d'une « extrême modernité » dans la construction de l'échantillonnage sociologique, en dépassant les catégories INSEE, utilisées initialement mais rapidement abandonnées au profit d'une nouvelle classification qui anticipait largement les reformulations qu'allait être proposées par Bourdieu, avec notamment la notion de capital culturel, infléchissant ainsi une sociologie purement économiste au profit d'une prise en compte parallèle des habitudes culturelles des témoins. Clairement identifiable au niveau du questionnaire, la filiation avec les conceptions qui se développaient au Centre de Sociologie Européenne dirigée à l'époque par Bourdieu au sein de l'EHESS, était explicitement revendiquée :

> « Le questionnaire sociolinguistique a été établi par les soins de M.B. Vernier, sociologue, élève du Professeur P. Bourdieu [...]. L'hypothèse sous-jacente à ce questionnaire prend appui sur une théorie sociale de la transmission de la culture et de la langue. L'origine sociale et le niveau d'éducation du témoin concourent, avec ses activités socio-professionnelles, à déterminer sa compétence linguistique d'une part, et son attitude envers les normes du langage d'autre part. Selon cette hypothèse, plusieurs types d'attitudes seraient en corrélation avec des niveaux socio-culturels, attitudes qui iraient de l'indifférence à la dévotion. Au sommet comme au bas de l'échelle socio-culturelle, cette attitude se traduirait par l'indifférence : mais tandis que dans la classe "Supérieure" cette indifférence est la marque d'une liberté par rapport à la norme et est liée à la capacité de jouer avec un grand nombre de registres, au bas de l'échelle, en revanche, cette indifférence est le résultat soit de l'ignorance, soit d'une certaine conscience de classe. Entre ces deux niveaux se situent d'une part ceux qui, ayant défini la norme linguistique, y adhèrent et cherchent à l'imposer ; d'autre part ceux qui, aspirant à une ascension sociale, s'efforcent d'imiter ce qui est pour eux la langue de prestige. » Blanc & Biggs (1971 : 18).

Le résultat fut la conception d'une nouvelle grille comprenant cinq catégories sociales – sténographiées « échelle AM » (du nom de son inventeur, Alix Mullineaux) – dont chacune était identifiée par une lettre de A à E, où l'on tente d'agréger aux CSP identifiées par l'INSEE des caractéristiques culturelles ainsi que des critères relatifs à la mobilité géographique.

L'adaptation nécessaire aux théories sociologiques contemporaines et aux nouvelles réalités sociétales n'a pas permis le maintien dans ESLO2 de la classification AM. C'est l'un des chantiers que l'équipe ESLO n'a pas pu rouvrir. Les locuteurs ont donc simplement été classés, conformément à la nouvelle nomenclature de l'INSEE, en 8 catégories socio-professionnelles[7].

Lors de la constitution du corpus ESLO-MD, afin de garantir une certaine comparabilité entre ESLO1 et ESLO2, nous avons repris, en cherchant à les adapter au cas par cas, les catégories AM. L'objectif était de traduire par l'un des agrégats A-E l'ensemble des informations dont nous disposons concernant le profil du locuteur, aussi bien au niveau des métadonnées (profession, études et diplômes, âge) qu'au niveau des informations restituables au sein de l'entretien concernant son appartenance sociale.

Menée sans appui sociologique et avec des critères complexes et non toujours explicités, la tâche était toutefois particulièrement fastidieuse et délicate, notamment concernant les catégories intermédiaires. En effet, il nous a

paru relativement aisé d'identifier des locuteurs comparables sur les « extrémités » de l'échelle, ainsi que nous pouvons le constater dans le tableau suivant :

Tableau 4. Exemple de profils ESLO2 et équivalence AM.

Âge 35-60						
Enregistrement	**Durée**	**Code locuteur**	**Sexe**	**Année**	**Cat. INSEE**	**Equivalence échelle AM**
ENT_1028	55	QF28	H	1952	Cadres/ profession intell. sup	A Gastro-entérologue
ENT_1056	56	TJ56	F	1974	Cadres/ profession intell. sup	A Psychiatre
ENT_1046	78	WZ46	H	1952	ouvrier	E Scolarité primaire
ENT_1085	56	RN 488	F	1961	employée	E Femme de ménage, brevet des collèges

Pour les catégories B-C-D, la classification a été nettement plus délicate. Quid par exemple de la femme au foyer, titulaire d'un bac + 5 en Droit, mariée, 49 ans, habitant Olivet – village aisé en périphérie d'Orléans – et parlant anglais, allemand et italien ?

Au-delà de ces problèmes inhérents à toute tentative de classification, il nous a paru que les difficultés proviennent surtout des nouvelles réalités sociales qui ont rendu nettement caduques les critères qui fondaient l'échelle AM. Le critère de diplôme, par exemple, ne pourrait avoir qu'une pertinence toute relative : un bac obtenu en 2015 ne peut avoir ni les mêmes débouchés professionnels ni la même valeur sociale qu'un bac de 1970. La féminisation de certains métiers, le déclassement de certains autres, l'émergence de nouvelles élites, l'apparition d'une nouvelle classe populaire d'origine immigrée qui semble cumuler les handicaps économiques et culturels, etc. sont autant de facteurs qui rendent aujourd'hui nécessaire une nouvelle nomenclature sociale critique. De même, les locuteurs classés tout en bas de l'échelle AM dans ESLO1 ne sont pas réellement comparables aux E de ESLO2. On pourrait d'ailleurs penser que les vrais E de notre époque (chômeurs n'ayant jamais cotisé ou très peu, jeunes déscolarisés, etc.) ne sont même pas présents dans ESLO2, notamment dans la catégorie des 15-35 ans. ESLO2 n'aura donc pas trouvé de solution à ce problème classique des enquêtes de ce type, déjà identifié dès ESLO1 où les ouvriers (hormis les délégués syndicaux, habitués à la prise de parole en public) étaient sous-représentés.

Les catégories ne pouvant s'éclairer que paradigmatiquement les unes par rapport aux autres, ce qui présuppose une vision sociale globale, on ne pouvait donc gérer la comparabilité des profils sociologiques qu'au cas par cas, armés de quelques décisions pragmatiques. Par exemple, on a cherché, à l'intérieur de chaque CSP, les profils prototypiques, en évitant, le plus possible, les cas hybrides et intermédiaires…

Toutes ces difficultés montrent que le sous-corpus constitué ne peut en aucun cas être considéré comme sociologiquement représentatif : il s'agissait simplement de réduire les variables entre les deux extraits ESLO1/ESLO2 en sélectionnant, dans deux corpus différents, des profils pouvant offrir la meilleure comparabilité qualitative possible :

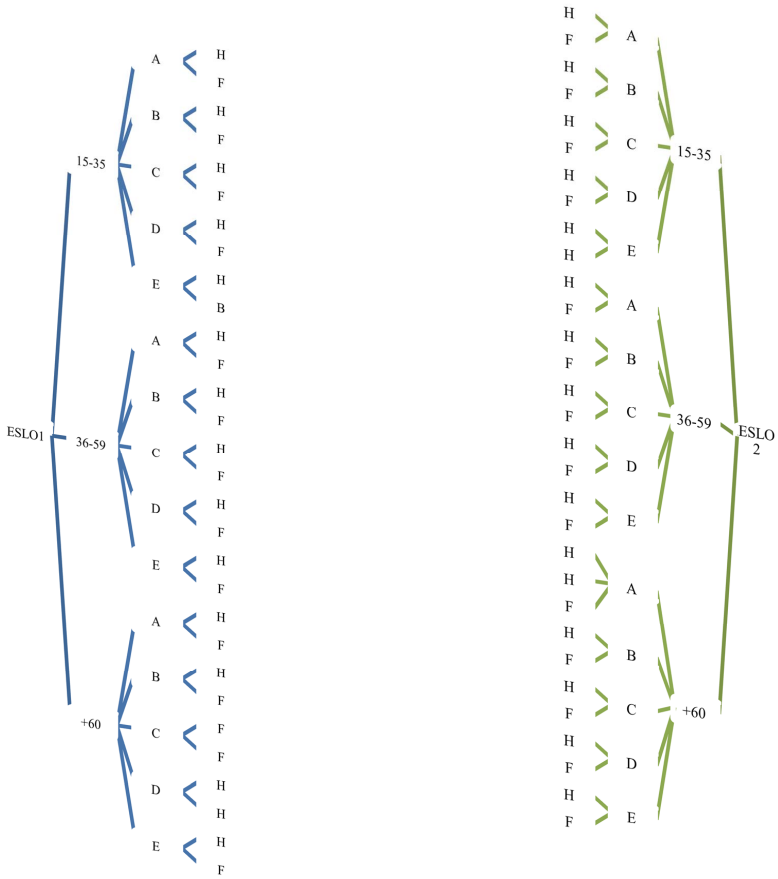

Fig. 1. Sélection de deux sous-corpus comparables selon variables Âge/CSP/Sexe.

3. Un corpus échantillonné et annoté pour la communauté

3.1. Outil et annotations

La constitution du sous-corpus micro-diachronique telle qu'elle a été présentée ci-dessus visait à mettre en lumière différentes formes de variation et à identifier des changements éventuels dans certaines zones « instables » du système en français hexagonal contemporain. Pour ce faire, nous avons mis en œuvre une procédure d'observation reposant sur un enrichissement du corpus en plusieurs étapes. La première étape d'enrichissement, automatique, a été réalisée avec le logiciel d'analyse textométrique TXM (Heiden & al. 2010)[8], doté d'un lemmatiseur et d'un étiqueteur en *part of speech* (Treetagger). C'est en balayant le corpus dans le concordancier par des requêtes combinées sur les lemmes, les étiquettes *pos* et les chaînes de caractères (via le système d'interrogation de TXM basé sur des expressions régulières CQL Corpus Query Langage) que nous avons pu extraire un premier ensemble de données, nous permettant de dégager des tendances générales quantitativement significatives pour l'objet observé. Cette première étape a été suivie d'une étape d'annotation sémantique fine, réalisée manuellement avec un étiquetage au cas par cas prenant en compte, grâce à la consultation systématique de la transcription alignée au son, le contexte interactionnel aussi large que nécessaire

à la bonne compréhension de l'emploi considéré. L'annotation sémantique, menée conjointement par deux annotateurs experts, s'est faite de manière concertée et inductive (en d'autres termes, dans une démarche *bottom-up*) pour faire émerger de l'observation les catégories pertinentes[9]. C'est en faisant suivre cette première phase d'annotation manuelle individuelle d'une phase de discussion et de négociation des annotations entre les deux annotateurs que les catégories ont été dégagées et stabilisées progressivement[10]. Nous avons finalement élaboré un guide comprenant une typologie d'exemples et un schéma arborescent du système d'annotation des types d'emplois. Ces annotations manuelles ont ensuite pu être réinjectées sous TXM et faire l'objet de nouvelles requêtes pour l'analyse et la quantification des phénomènes dégagés[11]. D'un point de vue scientifique, cette démarche présente l'avantage, en dépit du temps qu'exige une annotation minutieuse, de reposer sur un va-et-vient entre une analyse qualitative fine dans l'esprit des études sur l'oral en interaction, et une analyse quantitative, par la collecte systématique de données en nombre important.

3.2. D'autres études sur ESLO-MD

Après la publication de l'étude initiale sur ESLO-MD, d'autres collaborateurs plus ou moins proches de l'équipe ont souhaité à leur tour exploiter cette collection pour des objets parfois très différents de celui visé initialement (marqueurs discursifs, lexique, mode et temps, etc.[12]). A cet égard, il est intéressant de constater que la pertinence de la coupe micro-diachronique et de l'intervalle temporel considéré dépend de l'objet observé et ne permet pas toujours de statuer sur un changement en cours.

Quoi qu'il en soit, l'empan diachronique saisi par ESLO-MD, conjugué au caractère situé des données, fournit un angle d'observation de la variation et du changement particulièrement intéressant, notamment pour certains phénomènes, comme les marqueurs discursifs, le lexique, ou certaines formes verbales. De fait, lorsqu'elle est pertinente, la micro-diachronie ne montre pas toujours le même processus : grammaticalisation pour les formes verbales, pragmaticalisation pour les marqueurs discursifs, ou effet de mode, en particulier pour le lexique. Autrement dit, la description des usages peut s'inscrire plus ou moins pleinement dans la saisie du changement linguistique.

3.3. Des données accessibles en ligne

Compte tenu de l'intérêt porté à ce corpus par d'autres linguistes travaillant sur d'autres objets linguistiques, nous avons dû nous pencher sur la question du partage des données, en établissant un état des lieux des différentes solutions techniques existantes. La mise à disposition en ligne s'est tout de suite imposée comme une solution praticable dans le contexte scientifique actuel. Néanmoins, cette question s'est avérée complexe, non seulement pour des raisons techniques que nous exposons plus bas, mais aussi parce que le fait de rendre un corpus accessible pour la communauté des chercheurs oblige à s'interroger sur les objectifs scientifiques des futurs utilisateurs. La mise en ligne, qui doit parallèlement garantir un archivage pérenne et une protection de données parfois sensibles, doit anticiper des besoins différents, qui nécessitent des solutions techniques différentes, qu'il s'agisse de stocker et sauvegarder des données dans un espace de travail pour l'utilisation individuelle, de rendre accessible corpus et annotations à d'autres linguistes, pour faciliter par la transparence des sources l'évaluation des pairs en permettant la vérification-falsification, ou encore de rendre disponibles données primaires et secondaires pour la communauté scientifique dans une approche cumulative pour une réexploitation par la collectivité sur d'éventuels autres objets scientifiques. La plus-value scientifique de cette démarche est certaine : outre le fait d'inaugurer de nouvelles procédures vérificationnelles permettant d'augmenter la transparence et de favoriser l'échange scientifique, elle correspond aux pratiques les plus actuelles de partage, préconisées dans le cadre de la réflexion autour des humanités numériques notamment.

En fonction de ces cas de figure, plusieurs types de procédures peuvent être mises en place, depuis la simple référence et la description du corpus et des outils utilisés sur le site du projet[13] au dépôt des données dans des bases gérées par les infrastructures de recherche nationales pour un archivage plus ou moins pérenne (BnF, CoCoON[14], Ortolang[15], etc.). Cependant, en ce qui concerne le dépôt des données, plusieurs contraintes liées aux normes d'archivage existent. Dans CoCoON, par exemple, les données déposées sous forme de collection

font l'objet d'une notice et reçoivent un identifiant pérenne et un seul, attribué à l'ensemble de la collection, en l'occurrence l'ensemble de la collection ESLO1 et ESLO2 (données audio et transcriptions vérifiées et anonymisées). Pour éviter de déposer les données plusieurs fois ou de modifier l'identifiant initialement attribué, le dépôt dans CoCoON d'une sous-collection à l'intérieur d'une collection déjà déposée n'est pas possible. Ainsi, si la documentation partagée du travail de constitution de sous-corpus en tant que construction scientifiquement raisonnée est tout à fait possible via certains espaces de partage, sa pérennisation ne l'est pas encore.

Une autre contrainte vient d'un problème de versionnage des données secondaires, liées aux délais nécessaires à la chaîne de traitement, et d'une manière plus générale à la temporalité de l'activité scientifique. Ainsi, à l'époque de la constitution de la sous-collection ESLO-MD, toutes les transcriptions n'étaient pas prêtes pour la publication, tout en étant utilisables néanmoins pour des travaux de recherche diffusant uniquement des résultats et non des extraits du corpus. Ainsi, ESLO-MD est constitué en partie de versions antérieures à celles déposées ultérieurement pour ESLO1 et 2 en archivage pérenne (versions antérieures de transcription à l'orthographe non vérifiée et anonymisation partielle, nomenclatures différentes, etc.). Ce qui revient à dire qu'une partie du corpus ESLO-MD est publiable en accès libre, alors qu'une autre partie ne l'est pas, et sera donc en accès restreint (accès soumis à la signature d'une convention).

Malgré ces restrictions, des solutions s'offrent à nous pour favoriser le partage, grâce aux infrastructures créées pour développer le domaine des humanités numériques. Pour le cas particulier d'ESLO-MD, elles nous semblent être au nombre de trois. Nous présentons d'abord les deux que nous avons adoptées, en finissant par une troisième piste à envisager.

Sur le site ESLO (http://eslo.huma-num.fr/) associé à la plateforme de consultation de la base de données, il a été aisé de documenter le sous-corpus[16], en déposant un inventaire structuré de la sous-collection ESLO-MD renvoyant aux enregistrements et transcriptions, avec un accès restreint pour les versions non publiables (versions A), et en joignant un descriptif des études réalisées, les fichiers d'annotation et le guide (typologie d'exemples et arborescence). Cette opération, simple et rapide, ne permet pas cependant de référencement pérenne du travail.

Ce dernier est en revanche possible sur la plateforme ORTOLANG (https://www.ortolang.fr/), qui présente l'avantage, outre d'héberger des corpus et l'ensemble des données qui y sont associées, de leur attribuer un identifiant reconnu dans la communauté scientifique. Il est possible, dans cet espace, tout en renvoyant à l'outil utilisé pour son analyse (http://textometrie.ens-lyon.fr/), de déposer un sous-corpus de données primaires et secondaires assorti de ses métadonnées, ainsi qu'un fichier de corpus annoté généré par TXM (une collection TXM) et importable directement dans l'outil. Il est ainsi possible, pour tout utilisateur ayant installé TXM, d'importer l'intégralité du sous-corpus en une opération et de procéder à des manipulations sur les annotations. On adjoint à cela un document descriptif explicitant la démarche de constitution du sous-corpus ainsi que le guide d'annotation comportant les requêtes CQL pour une plus grande maniabilité et transparence[17].

L'inconvénient des solutions ci-dessus réside dans l'accès à l'outil et sa manipulation. Ainsi, pour interroger le corpus dans le cadre d'une procédure de vérification par exemple, le chercheur intéressé devra installer l'outil et utiliser une interface dont la manipulation nécessite souvent une assistance ingénieure, ce qui risque de freiner largement la diffusion effective du corpus enrichi. Une dernière piste resterait alors à étudier, celle du dépôt, nécessairement en accès restreint, sur la version web de l'outil TXM, actuellement en développement[18]. À l'aide du tutoriel mis en ligne par les auteurs (cf. Heiden, Magué & Pincemain 2010), ainsi qu'un guide spécifique à nos données, la tâche serait alors plus aisée.

Conclusion

Un examen minutieux de certains travaux antérieurs montre clairement que les analyses sont directement tributaires des données examinées (à la fois sur le plan quantitatif et qualitatif) et du prétraitement qu'ils ont dû subir (notamment lors de l'annotation, i.e. les catégories retenues et les valeurs qui leur sont attribuées), ce qui explique selon nous l'essentiel des divergences constatées entre études descriptives portant sur le même objet. Afin de garantir la falsifiabilité des résultats, il semble désormais incontournable d'assurer un accès aux données primaires et secondaires. Cela permettrait parallèlement de rentabiliser collectivement la démarche, réputée

chronophage et donc couteuse, de constitution d'un corpus et de son enrichissement. Après la diffusion, à une large échelle de différentes bases de données, y compris orales, c'est l'ère de diffusion de corpus d'étude, ayant concrètement donné naissance à des analyses précises, qui s'ouvre, pour permettre à la fois la vérification par les pairs des analyses déjà entreprises, et encourager de nouvelles études sur des objets différents. On pourrait également penser que puissent naitre des analyses sur un objet déjà étudié à partir du même corpus, mais avec un enrichissement différent, voire des études sur un même objet sur un corpus différent (avec reprise ou non de l'annotation préalablement utilisée) lorsque la démarche d'explicitation aurait montré à d'autres chercheurs l'inadéquation ou les limites du corpus initial par exemple.

Le seul obstacle à une telle perspective consiste en l'absence, à une époque en pleine évolution mais où les linguistes n'ont pas encore tous une expertise ou un soutien d'ingénierie, d'outils de TAL clé en main, pouvant aisément faire la jonction entre données primaires et secondaires d'une part, et analyse de l'autre.

En attendant l'apparition de tels outils – nous avons donné l'exemple de la version web de l'outil TXM, actuellement en développement – qui permettraient un accès total aux publications, aux annotations et aux données primaires, nous proposons avec ce cas pratique un aperçu de ce qu'il est possible de mettre en œuvre pour rendre un corpus, en l'occurrence un sous-corpus issu d'une collection plus grande, accessible et réutilisable par la communauté.

Références bibliographiques

Abouda, L. & Baude, O. (2007). Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des Eslo », in F. Rastier et M. Ballabriga (dir.), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Actes du XXVIIe Colloque d'Albi « Langages et Signification » : 161-168.

Abouda, L. & Skrovec, M. (2015). Du rapport entre formes synthétique et analytique du futur. Étude de la variable modale dans un corpus oral micro-diachronique, *Revue de Sémantique et Pragmatique*, n° 38, 35-57.

Abouda, L. & Skrovec, M. (2017a). Du rapport micro-diachronique futur simple/futur périphrastique en français moderne. Étude des variables temporelles et aspectuelles. *Corela* [En ligne], HS-21 | 2017, mis en ligne le 30 janvier 2017, consulté le 20 février 2017. URL : http://corela.revues.org/4804

Abouda, L. & Skrovec, M. (2017b). Alternance futur simple /futur périphrastique : variation et changement en français oral hexagonal. *Revue de Sémantique et Pragmatique*, 41-42 : 155-179.

Bergounioux, G., Baraduc, J. & Dumont, C. (1992). L'Etude sociolinguistique sur Orléans (1966-1991), 25 ans d'histoire d'un corpus. *Langue française*, 93, 74-93.

Blanc, M. & Biggs, P. (1971). L'enquête socio-linguistique sur le français parlé à Orléans, *Le Français dans le Monde*, 85, 16-25.

Mullineaux, L.A. & Blanc, M. H. A. (1982). The problem of classifying the population sample in the socio linguistic survey of Orléans (1969) in terms of socio-economic, social and educational categories. *Review of Applied Linguistics*, 55, 3-37.

Blondeau, H., Sankoff, G. et Charity, A. (2002). Parcours individuels dans deux changements linguistiques en cours en français montréalais. *Revue québécoise de linguistique*, Volume 31, numéro 1, 13-38.

Bourdieu, P. (1979). *La Distinction : critique sociale du jugement*. Paris, Minuit.

Fleury, S. & Branca, S. (2010). Une expérience de collaboration entre linguiste et spécialiste de TAL : L'exploitation du corpus CFPP 2000 en vue d'un travail sur l'alternance Futur simple / Futur périphrastique, *Cahiers AFLS*, Volume 16(1).

Habert, B. (2000). Des corpus représentatifs : de quoi, pour quoi, comment ?, *in* M. Bilger (éd.), *Linguistique sur corpus. Études et réflexions*, Perpignan, Presses Universitaires de Perpignan, 11-58. Version en ligne : http://icar.univ-lyon2.fr/ecole_thematique/idocora/documents/Habert_des_corpus_representatifs.pdf

Habert, B. (2005). *Instruments et ressources électroniques pour le français*. Paris/Gap, Ophrys, col. « L'Essentiel Français ».

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris, Armand Colin et Masson.

Heiden, S., Mague, J-Ph., Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Sergio Bolasco, Isabella Chiari, Luca Giuliano. *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, Jun 2010, Rome, Italie. Edizioni Universitarie di Lettere Economia Diritto, 2 (3), pp.1021-1032, 2010. <halshs-00549779>

Gudrun, L. & Léglise, I. (2013). Variations et changements linguistiques. In Wharton S., Simonin J. *Sociolinguistique des langues en contact*, ENS Editions, pp.315-329.

Pumain, D. (2005). Cumulativité des connaissances. *Revue européenne des sciences sociales* [En ligne], XLIII-131 | 2005, mis en ligne le 04 novembre 2009, consulté le 22 décembre 2017. URL :http://journals.openedition.org/ress/357 ; DOI: 10.4000/ress.357

Rendulic, N. & Kanaan-Caillol, L. (2016). *Je crois que, je pense que* : valeurs et variation dans un corpus oral diachronique. Actes du 5e Congrès Mondial de Linguistique Française - CMLF2016**,** 4-8 juillet 2016, Université François Rabelais de Tours. SHS Web of Conferences, 27 (2016). DOI: https://doi.org/10.1051/shsconf/20162702014

Siouffi G., Wionet C. & Steuckardt A. (2012). Comment enquêter sur les diachronies courtes et contemporaines. *Congrès Mondial de Linguistique Française – CMLF 2012*, SHS Web of Conferences 1.

Svartvik, J. (1992). Corpus linguistics comes of age. In J. Sartvik (ed.) *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82*. Berlin & New York: Mouton de Gruyter, 7-16.

Thibault, P. & Vincent, D. (1990). *Un corpus de français parlé. Montréal 84 : Historique, méthodes et perspectives de recherche*, Québec, Université Laval.

---

[1] lotfi.abouda@univ-orleans.fr
[2] marie.skrovec@univ-orleans.fr
[3] Ce corpus est librement disponible en ligne, sur le site http://eslo.huma-num.fr
[4] Voir notamment Abouda & Skrovec (2015) et Abouda & Skrovec (2017a, b).
[5] Enregistrement d'une journée entière d'un locuteur-auditeur.
[6] Entretiens dans le Centre Médico Psycho Pédagogique entre un parent et une assistante sociale.
[7] L'INSEE retient les 8 CSP suivantes :
1 Agriculteurs exploitants
2 Artisans, commerçants et chefs d'entreprise
3 Cadres et professions intellectuelles supérieures
4 Professions Intermédiaires
5 Employés
6 Ouvriers
7 Retraités
8 Autres personnes sans activité professionnelle.
Voir : http://www.insee.fr/fr/methodes/?page=nomenclatures/pcs2003/pcs2003.htm
[8] http://textometrie.ens-lyon.fr/
[9] Voir Abouda & Skrovec (2015) et (2017a et b).
[10] Si l'un des deux rapporteurs anonymes de cet article estime qu'il aurait été intéressant de mesurer le taux d'accord inter-annotateurs avant la phase d'adjudication, ce qui aurait permis de juger de la difficulté de la tâche et aurait constitué un indice précieux pour des tâches d'annotations similaires, nous estimons que ce type de démarche aurait été surtout nécessaire dans une perspective d'annotation, sinon automatisable, au moins susceptible d'être poursuivie par des annotateurs non-experts. Or, vu la complexité de l'annotation sémantique, la perspective nous paraît pour le moins prématurée.
[11] A l'époque du travail d'annotation, comme il n'était pas encore possible de créer directement dans l'outil des étiquettes nouvelles assorties de propriétés, il fallait procéder à un export du fichier d'annotation, pour le modifier, avant de le réinjecter dans l'outil. Cette possibilité a néanmoins été implémentée dans une version plus récente de TXM.
[12] Voir notamment Rendulic & Kanaan-Caillol (2016).
[13] http://eslo.huma-num.fr/
[14] https://cocoon.huma-num.fr/exist/crdo/
[15] https://www.ortolang.fr/
[16] http://eslo.huma-num.fr/index.php/pagelarecherche/projets-de-l-equipe-et-sous-corpus/eslo-md
[17] Laboratoire Ligérien de Linguistique - UMR 7270 (LLL) (2018). *ESLO-MD : Enquêtes Socio-Linguistiques à Orléans : Corpus Micro-Diachronie* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/eslo-md.
[18] http://portal.textometrie.org/demo/